From Erewhon to AlphaGo: For the Sake of Human Dignity, Should We Destroy the Machines?

Roger Brownsword*

ABSTRACT

This paper asks whether, for the sake of human dignity, regulators should adopt a precautionary approach to the development of smart machines. Having identified a set of essential (or commons') conditions for the existence of human social agents, including respect for human dignity in both foundational and non-foundational senses, consideration is given to human reliance on personal digital assistants, to the development of autonomous vehicles and lethal autonomous weapons systems, and to the use of smart machines in the criminal justice system. The paper concludes that, while smart machines should not be destroyed, a degree of precaution for the sake of human dignity is warranted. In particular, it is recommended that international agencies should monitor the impact of smart machines on the commons' conditions; and that national commissions should facilitate the articulation of the local social licence for the development and application of such machines.

ARTICLE HISTORY Received January 5, 2017; Accepted February 12, 2017

KEYWORDS Smart machines; machine learning; human dignity; personal digital systems, autonomous vehicles; autonomous weapons; criminal justice; precaution

1. Introduction

This is a paper about the regulatory challenges and opportunities presented by today's 'smart' technologies—that is to say, those technologies that, enabled by machine-learning, have the capacity not only to operate in an intelligent manner but also, in many cases, to outperform humans.[1] In particular, it is a paper about the relationship between smart machines and human dignity.

Although smart machines are very much the talk of the twenty-first century, we can start in the second half of the nineteenth century, with Samuel Butler's novel, *Erewhon*.[2] To Butler's Victorian readers, the practices of the inhabitants of that eponymous distant land must have seemed quite extraordinary. How could the Erewhonians think it appropriate to punish those

* **CONTACT** Roger Brownsword, roger.brownsword@kcl.ac.uk, Dickson Poon School of Law, King's College London, Strand, London WC2R 2LS.

1 For a very clear account of the way in which 'machine learning' operates and its potential utility in legal practice (particularly in the context of litigation), see Harry Surden, 'Machine Learning and Law' (2014) 89 *Washington Law Review* 87.

2 First published 1872. Available at www.planetebook.com (last accessed February 3, 2017).

who fall ill while sympathising with those who commit crimes? How could they think it rational to destroy their machines?[3] How could such intelligent and technologically sophisticated people have gone backwards in this way?

Yet, the beauty of *Erewhon* is that, to some present-day readers—particularly readers who are familiar with, say, *Superintelligence*,[4] *Homo Deus*,[5] *Here be Dragons*,[6] or *A Dangerous Master*[7]—the practices of the Erewhonians might seem to be anything but benighted. For example, were the Erewhonians so stupid in supposing that, where an individual misbehaves, such conduct is to be treated as 'the result of either pre-natal or post-natal misfortune'?[8] Is it so ridiculous to think that, with the acceleration in technological development, machines might become much smaller and smarter, capable of reproducing themselves, communicating with one another, and displaying various degrees of intelligence (if not intelligence as humans understand it) and agency? Most importantly, would it be crazy to regard machines as a threat to the human condition that warranted at least some precautionary measures—albeit perhaps not precaution on the scale exercised by the Erewhonians who destroyed 'all the inventions that had been discovered for the preceding 271 years'?[9]

As is well-known, human chess players no longer reign supreme. Indeed, it is already 20 years since Deep Blue was programmed to beat Garry Kasparov the then world champion chess player;[10] and, since that time, the processing power of computers has continued to grow in the way predicted by Moore's Law.[11] However, we are still coming to terms with the capacity of machines to learn how to get better at playing games that are even more computationally challenging than chess. Notably, in March 2016, when AlphaGo defeated Lee Sedol, the South Korean world champion Go player, even the developers of this smart technology were surprised by the effectiveness of their machine-learning (ML) algorithms. Recalling Butler's cautionary tale, the focal question in this paper is whether we are at a point

---

3 Butler (n 2), Chs XXIII-XXV.

4 Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

5 Yuval Noah Harari, *Homo Deus* (Harvill Secker, 2016) for a possible future in which smart machines render humans largely surplus to requirement. As Harari puts it, when there are IBM Watsons around, 'there is not much need for Sherlocks' (at 316).

6 Olle Häggström, *Here be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press, 2016).

7 Wendell Wallach, *A Dangerous Master* (Basic Books, 2015) esp. Ch 12.

8 Butler (n 2) p 96. Compare, e.g., Tom Whipple, 'Three-year-olds can be identified as criminals of the future' *The Times*, December 13, 2016, p 1 (reporting that a test which rated a child's IQ and self-control together with information about deprivation and maltreatment, was highly predictive in relation to the future criminality, poor health and unemployment of a particular group).

9 Butler (n 2) p 260.

10 https://en.wikipedia.org/wiki/Deep_Blue_versus_Garry_Kasparov (last accessed, February 3, 2017).

11 https://en.wikipedia.org/wiki/Moore%27s_law (last accessed, February 3, 2017).

when we should suspend further development of smart machines. More specifically, the question is whether we should so act *for the sake of human dignity*—a notion not explicitly part of the Erewhonians' thinking but which we might take nowadays to be central to any rational and justified limits that are set on the development and application of modern technologies.[12]

While the development of smart *games-playing* machines (such as IBM's Watson[13] and AlphaGo itself) might have some impact on human games-players—indeed, some might argue that there is an 'indignity' in a human being beaten by a machine—I suggest that this raises no serious questions about human dignity. However, there are other applications of ML where such questions might be raised. From the many potential applications of machine learning, I propose to consider four particular cases, namely: (i) the use of ML in consumer recommender systems together with personal digital assistants (or cyberbutlers); (ii) the programming of moral decision-making into ML-enabled autonomous vehicles; (iii) the use of ML in lethal autonomous weapons systems; and (iv) the use of ML for the purposes of profiling, prediction, and prevention in the criminal justice system.

The paper is in five principal parts. First, starting with the idea that what we humans have in common is our 'humanity', I sketch a triple bottom line for twenty-first century regulators whose responsibilities are to protect and preserve (including by taking precautionary measures) the essential conditions (the commons' conditions) for a community of human agents with moral aspirations. These conditions relate to the possibility of human existence, the possibility of agency and self-development, and the possibility of moral development and moral action. While familiar theories of human dignity—notably, those liberal rights-based and conservative duty-based theories that are so often appealed to in order to contest such matters as the treatment of human embryos, the commercialisation and commodification of the human body, the permissibility of assisted suicide and euthanasia, and so on[14]—articulate in ways that underline these (commons') responsibilities of both regulators and regulatees, I suggest that there is a more formal foundational sense in which human dignity is at stake. Quite simply, this is the idea that, because humans express their dignity when they freely choose to do the right thing, one aspect of the commons' responsibilities is to maintain the conditions for the development and operationalisation of human dignity so understood. While this understanding of human dignity does not prescribe what prospective moral agents should

---

12 Compare Marcus Duwell, 'Human Dignity and the Ethics and Regulation of Technology' in Roger Brownsword, Eloise Scotford, and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation and Technology* (Oxford University Press, 2017) (forthcoming) who remarks that 'to investigate the relationship between human dignity and the regulation of technologies is about nothing less than the question of what an appropriate normative framework for the contemporary technology-driven world could be.'

13 On which, see, Jason Millar and Ian Kerr, 'Delegation, relinquishment, and responsibility: The prospect of expert robots' in Ryan Calo, A. Michael Froomkin, and Ian Kerr (eds), *Robot Law* (Elgar, 2016) 102-127.

14 See Roger Brownsword, 'Human Dignity from a Legal Perspective', in M.Duwell, J. Braavig, R. Brownsword, and D. Mieth (eds), *Cambridge Handbook of Human Dignity* (Cambridge University Press, 2014) 1-22.

judge to be right and what wrong—it is for each agent to make their own judgment as to what, in the instant case, is the right thing to do—it does prescribe the pre-conditions for agents to be able to reason and act in this way.[15]

Secondly, turning to applications of ML, I assess whether personal digital assistants (PDAs), cyberbutlers, and consumer recommender systems, and the like, might compromise human dignity—whether in the formal foundational sense or as understood in various liberal and conservative views. Thirdly, although the contexts are very different, I consider (i) whether the programming of moral decision-making into autonomous road traffic vehicles might compromise human dignity and (ii) the argument that the use of autonomous weapons is a violation of human dignity. In both cases, it might be argued that, where life-and-death choices or decisions are being made, humans should be 'in the loop'.[16] Fourthly, I consider whether the use of ML for the purposes of profiling, prediction, and prevention in the criminal justice system might compromise human dignity—again, whether in the formal foundational sense or as understood in various liberal and conservative views. Finally, I review the kinds of precautionary measures that might be taken, and how they might be justified, in response to the uncertain impact of ML. One of the key points here is that the arguments for precaution are altogether more compelling where the perceived threat of ML relates to the commons' conditions themselves (including the conditions that pertain to human dignity in the formal foundational sense) rather than to acts, activities, or practices that presuppose the existence of those conditions.

My conclusion, in company with the Erewhonians, is that there are reasons to be concerned about the machines. To be sure, smart cities and comprehensively intelligent machines might be some way off. Nevertheless, a culture of reliance on smart machines—whether PDAs or autonomous vehicles—might compromise the conditions for the moral development of human agents (and, with that, human dignity in a foundational sense); the use of ML for risk-assessment purposes in the criminal justice system poses obvious threats to a liberal conception of human dignity that attaches importance to due process and the rights of individuals; the use of lethal autonomous weapons might indirectly diminish respect for human dignity and the right to life; and, where ML is implicated in the technological management of risk, there are again concerns about how this impacts on the pre-conditions for moral community (especially, on human dignity in the foundational sense). However, by contrast with the Erewhonians, I do not propose that we should terminate the development of machine learning or destroy all records of machine-related learning since the Industrial Revolution. Rather, I suggest that we need to work on two things. First, we need to ensure that we have in place international agencies (new or existing) whose responsibility is to

---

15 Compare Roger Brownsword, 'Human Dignity, Human Rights, and Simply Trying to Do the Right Thing' in Christopher McCrudden (ed), *Understanding Human Dignity* (Proceedings of the British Academy 192) (The British Academy and Oxford University Press, 2013) 345-358.

16 Compare, e,g, Adam Saxton, '(Un)Dignified Killer Robots? The Problem with the Human Dignity Argument' (Lawfare Institute) available at https://www.lawfareblog.com/undignified-killer-robots-problem-human-dignity-argument (last accessed October 29, 2016).

monitor the impact of machine learning on the commons' conditions and to take whatever precautionary measures might be required. Secondly, at national level, we need to establish Commissions whose responsibility, on the one hand, is to liaise with the international agencies and, on the other, to assist their communities in developing an informed and inclusive 'social licence'[17] for new technologies—that is to say, in the present context, the development of agreed terms and conditions for the socially acceptable use and application of machine learning in their territory.[18]

2. Humanity and the Triple Bottom Line for Regulators

According to Mary Aiken,

> the gender battles of the previous century will seem like a picnic compared with what's coming next: the battle between humans and artificial intelligence. It's time to forget about our differences—gender, ethnicity, nationality—and focus on the thing that unites us, our humanity.[19]

While we might not share Aiken's view that what lies ahead is a 'battle' between humans and smart machines,[20] it seems to me that her call for humans to come together under the banner of their shared humanity is exactly right.

Guided by Aiken, we might think that a plausible response to the question in this paper—that is to say, the question of whether ML might compromise human dignity—runs along the following lines. It is for each society to debate and then determine the terms and conditions on which ML is to be licensed. In some communities, there might be no licence at all or, as in Erewhon, a licence once granted might be revoked; but, in other communities, the use and

---

17 Here, I am using the term 'social licence' in a sense that is akin to a 'social contract'. This is broader than saying that some particular x (whether an act, practice, or policy), even though 'compliant' relative to some positive standards (e.g., legal or regulatory standards, professional standards, or the like) nevertheless lacks a social licence. That said, there is interesting work to be done in exploring the gap between technical rule-compliance and social acceptability (for example, in relation to tax avoidance schemes, corporate social responsibility, parliamentarians' expenses, and so on). For an example of this gap, see Pam Carter, Graeme T. Laurie, and Mary Dixon-Woods, 'The social licence for research: why *care.data* ran into trouble' *Journal of Medical Ethics* (published online 23 January, 2015) doi. 10.1136/me3dethics-2014-102374.

18 Compare Geoff Mulgan's proposal for the establishment of a Machine Intelligence Commission: available at http://www.nesta.org.uk/blog/machine-intelligence-commission-uk (blog 'A machine intelligence commission for the UK', February 22, 2016: last accessed December 11, 2016); Olly Bustom et al, *An Intelligent Future? Maximising the Opportunities and Minimising the Risks of Artificial Intelligence in the UK* (Future Advocacy, London, October 2016) (proposing a Standing Commission on AI to examine the social, ethical, and legal implications of recent and potential developments in AI); HC Science and Technology Committee, *Robotics and Artificial Intelligence* HC 145 2016-17.

19 Mary Aiken, *The Cyber Effect* (John Murray, 2016) at 316.

20 If we conceive of a 'battle' in terms of a conflict between groups of agents with a degree of 'consciousness' and 'intentionality', then this raises the question of whether smart machines could ever cross such a threshold. For reflections on such matters, see Amedeo Santosuosso, 'The Human Rights of Nonhuman Artificial Entities: An Oxymoron?' (2014) 19 *Jahrbuch für Wissenschaft und Ethik* 203.

application of ML will be authorised and limited by a social licence. The extent to which these licensing decisions reflect a concern about human dignity is likely to vary from one society to another, depending not least on the particular conception of human dignity that prevails in each society. Nevertheless, if we follow Aiken in focusing on our humanity, I suggest that we might find some significant common ground.

In this section of the paper, I will start by sketching the three dimensions of this common ground—these dimensions relating respectively to the existence pre-conditions, the self-development and agency pre-conditions, and the moral development and opportunity pre-conditions for humanity; then, I will explain how I see human dignity fitting into, and relating to, these pre-conditions; and, finally, I will indicate a puzzle about the scope and application of the dignity conditions that I will have to leave for another time.

2.1 *Common ground*

In what respect might there be common ground between humans and their communities? If we equate 'humanity' with a community of human agents with moral aspirations, then what should we take to be the common ground between such agents? Or, to put this slightly differently, what common ground should be assumed by regulators who are charged (among other things) with sustaining the conditions for such communities?[21] At once, it seems to me, we have a triple bottom line for today's regulators. Quite simply, the terms and conditions of any regulatory (or social) licence for new technologies should be such as to protect, preserve, and promote:

- the essential conditions for human existence (given human biological needs);
- the generic conditions for human agency; and,
- the essential conditions for the development and practice of moral agency.

Moreover, these are imperatives for regulators in all regulatory spaces, whether international or national, public or private. From the responsibilities for the commons, there are no exemptions or exceptions; we are dealing here with principles that are truly cosmopolitan.[22]

In the first instance, regulators should take steps to protect, preserve and promote the natural ecosystem for human life. Starting with the maintenance of the so-called 'planetary boundaries',[23] regulators need to prevent the occurrence of (or, at any rate, minimise the

---

[21] I am grateful to Christian Illies for pressing me to be more precise about the viewpoint from which the common ground implicit in humanity is to be developed. While there might be some views that purport to challenge the need to preserve the bottom-line conditions—even the view that humanity is morally required to put an end to its own existence—these are either views that regulators need not take seriously or views that can only be presented to regulators if the bottom-line conditions are in place. Alternatively, following a suggestion by Illies, the argument might be strengthened if a further set of essential conditions, for the long-term sustainability of the moral community, were to be added.

[22] Compare Roger Brownsword, *Rights, Regulation and the Technological Revolution* (Oxford University Press, 2008) Ch. 7; and Regulatory Cosmopolitanism: Clubs, Commons, and Questions of Coherence" TILT Working Papers No 18 (2010).

[23] Understood as 'the non-negotiable planetary conditions that humanity needs to respect in order to avoid the risk of deleterious or even catastrophic environmental change at continental to global scales': see, J.

damage caused by) human-initiated existential threats—for example, the threats presented by ozone-depleting chemicals, dangerous pathogens, the proliferation of nuclear weapons, (arguably) large particle accelerators and colliders,[24] and so on. Secondly, the conditions for meaningful self-development and agency need to be constructed (largely in the form of positive support and negative restriction): there needs to be sufficient trust and confidence in one's fellow agents, together with sufficient predictability to plan, so as to operate in a way that is interactive and purposeful rather than merely defensive. Fear, as Robert Nozick famously highlighted, can have a corrosive impact on agency.[25] Thirdly, there need to be conditions for the moral development of agents as well as for the practising of moral agency—that is to say, the moral segment of the commons features both developmental and opportunity conditions.

Now, the point about these three bottom lines is that they are conceived of as being pre-competitive or pre-conflictual; the tensions, competing demands, purposes and priorities that characterise much social life are all to come. These bottom line 'pre-conditions' (whatever they are agreed to be) are, by definition, neutral as between one human and another, between one agent and another, and between one agent with moral aspirations and another (or, between one moral viewpoint and another).[26] These are pre-conditions that represent a 'commons' that reflects the needs of all humans, irrespective of their particular projects and plans as agents, and irrespective of their particular moral beliefs. If, during the course of deliberative democratic debate, anyone proposes that smart machines should be licensed to operate in ways that might compromise any of these bottom-line conditions, regulators should treat such a proposal as wholly 'unreasonable'. Of course, determining the nature of these conditions will not be a mechanical process and I do not assume that it will be without its points of controversy.[27] Nevertheless, let me give an indication of how I would understand the distinctive contribution of each segment of the commons and then explain how I see human dignity fitting into the picture.

First, the commons must secure the essential conditions for *human* existence—that is, for an ecosystem that is capable of supporting human life. At minimum, this entails that the physical well-being of humans must be secured; humans need oxygen, they need food and water, they need shelter, they need protection against contagious diseases, if they are sick they need

---

Rockström et al, 'Planetary Boundaries: Exploring the Safe Operating Space for Humanity' (2009) 14 *Ecology and Society* 32 (http://www.ecologyandsociety.org/vol14/iss2/art32/) (last accessed November 14, 2016). See, too, Kate Raworth's notion of 'doughnut' economics: http://www.kateraworth.com/doughnut/ (last accessed November 14, 2016).

24 See Wendell (n 7), 1-7 (for risks concerning the creation of black holes and strangelets).

25 Robert Nozick, *Anarchy, State, and Utopia* (Blackwell, 1974).

26 But, note my caveats in n 21.

27 Moreover, even if it is agreed where the bottom lines are to be drawn, a community still has to decide how to handle proposals for uses of smart machines that do not present a threat to any of the bottom line conditions.

whatever medical treatment is available, and they need to be protected against assaults by other humans or non-human beings. It follows that the intentional violation of such conditions is a crime against, not just the individual humans who are directly affected, but humanity itself.[28] If we limit the essential conditions for human existence to such survival resources, this is compatible with both a good deal of psychological distress as well as restrictions on human freedom (the former possibly stemming from the latter). If smart machines were to enslave humans, and possibly keep a few well-maintained human specimens as exhibits in a zoo, this would be a fundamental violation of human *agenc*y but not of the essential conditions for human existence. On the other hand, ML employed in weapons of mass destruction would be a clear case of a violation of these essential conditions.

Secondly, there are the generic conditions for human agency. Let me assume that the distinctive capacities of prospective agents include being able:

- to freely choose one's own ends, goals, purposes and so on ('to do one's own thing')
- to understand instrumental reason
- to prescribe rules (for oneself and for others) and to be guided by rules (set by oneself or by others)
- to form a sense of one's own identity ('to be one's own person').

Accordingly, the essential conditions are those that support the exercise of these capacities. With existence secured, and under the right conditions, human life becomes an opportunity for agents to be who they want to be, to have the projects that they want to have, to form the relationships that they want, to pursue the interests that they choose to have and so on. In the twenty-first century, no other view of human potential and aspiration is plausible; in the twenty-first century, it is axiomatic that humans are prospective agents and that agents need to be free.

The gist of these agency conditions is nicely expressed in a recent paper from the Royal Society and British Academy where, in a discussion of data governance and privacy, we read that:

> Future concerns will likely relate to the freedom and capacity to create conditions in which we can flourish as individuals; governance will determine the social, political, legal and moral infrastructure that gives each person a sphere of protection through which they can explore who they are, with whom they want to relate and how they want to understand themselves, free from intrusion or limitation of choice.[29]

---

[28] Compare Roger Brownsword, 'Crimes Against Humanity, Simple Crime, and Human Dignity' in Britta van Beers, Luigi Corrias, and Wouter Werner (eds), *Humanity across International Law and Biolaw* (Cambridge University Press, 2014) 87-114.

[29] The Royal Society and British Academy, *Connecting Debates on the Governance of Data and its Uses* (London, December 2016) 5.

In this light, we can readily appreciate that what is dystopian about George Orwell's *1984*[30] and Aldous Huxley's *Brave New World* [31]is not that human *existence* is compromised but that human *agency* is compromised.[32] We can appreciate, too, that today's dataveillance practices, as much as 1984's surveillance, 'may be doing less to deter destructive acts than [slowly to narrow] the range of tolerable thought and behaviour.'[33]

Thirdly, where human agents have moral aspirations, the commons must secure the conditions for a moral community. Agents who reason impartially will understand that each human agent is a stakeholder in the commons that protects the essential conditions for human existence together with the generic conditions of agency; and that these conditions must, therefore, be respected. Beyond these conditions, the moral aspiration is to do the right thing relative not simply to one's own interests but relative to the interests that other human agents might have. While respect for the commons' conditions is binding on all human agents, these conditions do not rule out the possibility of moral contestation and moral pluralism. Rather, these are pre-conditions for moral debate and discourse, giving each agent the opportunity to develop his or her own view of what is morally prohibited, permitted, or required in relation to those acts, activities and practices that are predicated on the existence of the commons.

2.2 *Human dignity*

How does human dignity enter into this picture? I suggest that it enters at two points—or, better, at two levels.

First, we can understand the idea of respect for human dignity as a demand that there is respect for the essential pre-conditions for the development and practice of moral agency (in other words, respect for the third of the bottom lines). As a shorthand, let us call this 'HD1'. Elaborating this notion, we would treat the distinctive value (or dignity) of humans as residing in their capacity to appreciate the importance of doing the right thing and then acting on that understanding. To the extent that human dignity is understood to be a virtue, it is the virtue of freely trying to do the right thing for the right reason. Because human dignity, so conceived, is impartial between agents who contest the criterion for doing the right thing, it sets a particular bottom line for any regulatory or social licence for machine learning. What is dystopian about Anthony Burgess' *A Clockwork Orange*[34], and what is at least worrying about the supposed utopia of B.F. Skinner's *Walden Two*[35], is that those who do the right

30 (Penguin Books, 1954) (first published 1949).

31 (Vintage Books, 2007) (first published 1932).

32 To be sure, there might be some doubt about whether the regulation of particular acts should be treated as a matter of the existence conditions or the agency conditions. For present purposes, however, resolving such a doubt is not a high priority. The important question is whether we are dealing with a bottom-line condition.

33 Frank Pasquale, *The Black Box Society* (Harvard University Press, 2015) at 52.

34 (Penguin Books, 1972) (first published 1962).

35 (Hackett Publishing Company Inc, reprinted 2005) (first published 1948).

thing—whether reformed criminals or young children—do not seem to have any choice in the matter.

Secondly, human dignity might enter the picture once again where it is put forward within moral discourse as the substantive criterion for doing the right thing. For present purposes, I will assume that such appeals to human dignity are either rights-based, the idea being that human dignity 'constitutes the real basis of fundamental rights'[36] (for short, we can call this 'HD2a'), or duty-based, the idea being that human dignity is the source of duties that humans owe not only to other humans but also to their communities and to themselves (for short, we can call this 'HD2b').

On this analysis, the general question of whether any applications of ML present a threat to human dignity translates into two more specific questions. First, will the application compromise the bottom line conditions for moral community (i.e., HD1)? Secondly, is the application compatible with human dignity as the particular standard of doing the right thing (i.e., HD2a or HD2b)? Whereas all human agents who have moral aspirations will accept that there can be no compromising of the former, their views in relation to the latter will depend upon how they understand the relationship between human dignity and the particular criterion of right action.[37]

It remains only to say that, while there are many questions on which proponents of HD2a differ from proponents of HD2b, they must agree that humans have a moral responsibility to respect the bottom-line conditions. Hence, if some act or activity or technological application compromises any part of these conditions, it necessarily will involve a violation of both HD2a and HD2b; and, if it is the conditions for moral development and opportunity that are compromised, then HD1 is also distinctively engaged.

2.3 *Unrestricted and restricted moralism*

This leaves a question that I will not pursue in this paper but which I want to flag up as needing further attention. The question is whether there might be some differences between moralists as to whether a technological fix, rather than a rule, may be legitimately employed to protect the existence and agency conditions.

Briefly, what I have in mind is a possible distinction between (i) 'unrestricted moralists' who hold that human agents should aspire to do the right thing for the right reason in relation to all

---

36 As the matter is expressed in the explanatory notes to Article 1 of the EU Charter of Fundamental Rights (see http://fra.europa.eu/en/charterpedia/article/1-human-dignity) (last accessed November 2, 2016).

37 Arguably, the generic conditions for agency already presuppose not only the conditions for moral development and opportunity but a particular set of substantive rights-based moral principles (along the lines of HD2a). Seminally, see Alan Gewirth, *Reason and Morality* (University of Chicago Press, 1978); and Deryck Beyleveld, *The Dialectical Necessity of Morality* (University of Chicago Press, 1991). However, for present purposes, so long as it is accepted that regulators are dealing with a context of moral aspiration, it is not critical to cash this argument.

acts, including acts that concern the commons' conditions themselves, and (ii) 'restricted moralists' who hold that this moral aspiration is *fully* engaged only in relation to acts that, while relying on the commons, do not directly impact on the commons' conditions themselves. Let me emphasise that those who take a restricted approach are not disputing that we have moral rights and responsibilities in relation to either the existence or the agency conditions, and that there are overwhelming moral reasons for prioritising the protection and preservation of the commons, indeed that these are core human responsibilities.

To illustrate the difference between these two degrees of moralism, let us assume that, in a particular community, the core rules of the criminal law are treated as instruments that are designed to secure the commons' (agency) conditions. Now, when modern technologies of surveillance and DNA-identification are adopted in support of these laws, unrestricted moralists will be concerned that the amplification of prudential reasons for compliance might result in the crowding out of moral reasons. When the use of technology is taken a stage further, such that 'non-compliance' is prevented and 'compliance' is technologically forced, agents no longer reason in terms of what they ought to do, neither prudentially nor morally. For unrestricted moralists, this is a move in the wrong direction; HD1 is compromised. However, for those who take a restricted view, a technological fix for the sake of the existence or agency pre-conditions is not necessarily a problem.[38]

Having drawn this distinction, two questions arise. First, is there any rational basis for the (unrestricted moralist) view that, even where rules do not work in protecting the commons' conditions and where technological management might fix the problem, the priority is still to try freely to do the right thing? Secondly, does the restricted view leave much significant space for agents to debate what is morally required and for agents to do the right thing for the right reason? These are large questions that I have to leave for another day.

3. Machine Learning, Consumer Recommender Systems, Personal Digital Assistants and Cyberbutlers

In the early days of e-commerce, in a prescient article,[39] Richard Ford anticipated the profiling of consumer preferences and the servicing of one's consumption needs by automated processes, or by something akin to a 'cyberbutler'. Imagining such a future, Ford foresaw that a consumer would sign over their paycheck to the cyberbutler who would hold it in trust for the consumer's benefit; then, guided by the consumer's profile, the cyberbutler would place appropriate orders so that, each day, the consumer would 'come home to a selection of healthy and nutritious groceries from webvan.com or a Paul Smith shirt from

---

38 For these changes in the regulatory registers, see Roger Brownsword, 'Lost in Translation: Legality, Regulatory Margins, and Technological Management' (2011) 26 *Berkeley Technology Law Journal* 1321. And, for the general question of whether we should try to eliminate the possibility of offending, compare Michael L. Rich, 'Should We Make Crime Impossible?' (2013) 36 *Harvard Journal of Law and Public Policy* 795.

39 Richard T. Ford, 'Save the Robots: Cyber Profiling and Your So-Called Life' (2000) 52 *Stanford Law Review* 1572.

boo.com or the latest Chemical Brothers CD from cdnow.com'.[40] However, Ford's cyberbutlers, like today's recommender systems, go one step beyond repeat ordering. Additionally, they 'will suggest books I'll enjoy reading, recordings I'll like to listen to, restaurants I'll be glad I tried, even if I wouldn't have chosen any of them on my own.'[41] While some of these suggestions will coincide with my likes, others will not; but, the more that my preferences are revealed and used to refine the profile, the better my ML recommender systems will become in tailoring their advice. If these systems constantly improve, is there anything to cause human agents concern? If the recommendations got worse (in the sense of being further from our preferences or simply inappropriate given our age, needs, and interests) they might be a nuisance; but, if they get better, almost to the point that the profilers seem to know me better than I know myself, should we worry? And, if so, is it the compromising of human dignity that should be the reason for our worry?

To this, my short answer is that there might be issues here about the ideal conditions for human agency to flourish[42] as well as for the welfare of consumers.[43] Moreover, for those moralists who subscribe to either HD2a or HD2b, there might be some particular uses of PDAs that give rise to dignitarian concern.[44] However, there is also a more systemic concern: this is that, while recommender systems promise to give agents more options of the kind that they generally like to have and, to this extent, expand agent choices, there might be some inhibition on changing one's preferences and, thus, modifying the kind of person that one wants to be. In other words, such systems raise the question of whether there is a risk that consumers might become trapped in their own (machine learning-assisted) personal 'echo-chambers', and whether this risk needs to be managed.[45] The thought is that those consumers who simply want to experiment or change their profiles might find it difficult to do so. The challenge for society is to find a way of balancing the preferences of those agents who are

---

40 Richard Ford (n 39) at 1578. By now, the cyberbutler would have changed to different suppliers of groceries, shirts, and CDs (webvan.com and boo.com being spectacular dot.com failures, and a declining cdnow.com being purchased by Amazon).

41 Richard Ford (n 39) at 1576.

42 Compare, e.g., Sherry Turkle, *Alone Together* (Basic Books, 2011).

43 See the excellent analysis in Ariel Ezrachi and Maurice E. Stucke, *Virtual Competition* (Harvard University Press, 2016).

44 For a recent story that might trouble those who subscribe to HD2b, see Mark Bridge, 'Forget real women: men turn to Siri for their sexual thrills' *The Times*, October 27, 2016, p. 3.

45 See, e.g., European Data Protection Supervisor, *Towards a New Digital Ethics* (Opinion 4/2015) 11 September, 2015, at 13:

> Profiles used to predict people's behaviour risk stigmatisation, reinforcing collective stereotypes, social and cultural segregation and exclusion, with such 'collective intelligence' subverting individual choice and equal opportunities. Such 'filter bubbles' or 'personal echo-chambers' could end up stifling the very creativity, innovation and freedoms of expression and association which have enabled digital technologies to flourish.

perfectly happy to receive recommendations that point them towards more of the same against the preferences of those who wish to try different things.

This is not, however, the end of the matter: for, as Mireille Hildebrandt has argued, human agents who use a PDA will find that they are living in 'an onli*f*e world' in which their cyberbutlers also operate as agents.[46] To aid our imagination of what it might be like to live in such a world, Hildebrandt introduces us to 'Diana', a sales representative for a large hotel conglomerate, whose life is organised by a PDA 'that is distributed between [Diana's] smart phone, the system running her smart house, the smart car, her ubiquitous computing office platform, while being on speaking terms with other systems, like those for traffic control and healthcare, commercial and governmental service providers, as well as monitoring systems for private and public safety and security'.[47]

At first blush, the use of such PDAs might seem to be unproblematic. If Diana chooses to rely on her PDA, how does this challenge human dignity? However, if it turns out that Diana's freedom to use her PDA is more apparent than real, her 'consent' might not be adequate to authorise the kind of exploitation of her personal data that might trouble advocates of HD2a. More importantly, we need to take a harder look at the way in which Diana relies on her PDA. In particular, what if Diana comes to rely on her PDA to analyse her moral dilemmas and determine how she should act? For those who belong to an aspirant moral community, it is axiomatic that each agent should develop a sense of what it is to do the right thing for the right reason, and try always to do just that. Human dignity involves more than merely acting in line with the right thing; the paradigmatic expression of the dignity of humans is in doing what an agent judges to be the right thing even where there is an opportunity to do the wrong thing. In the light of this, it is one thing for agents to use their PDA as a moral 'critical friend', quite another to rely habitually on the PDA's moral expertise or self-consciously to delegate moral decision-making to their PDA; and, similarly, it is one thing for an agent freely to comply with legal rules but quite another (in techno-managed environments) to have no practical option other than to comply with the constraints imposed by whatever technological measures have been adopted. Where agents no longer freely make and act on their moral judgments, we should question whether the conditions for moral community—and, concomitantly, for human dignity (as HD1)—are being compromised.[48]

In these short remarks, there are, at least, two invitations to engage precaution. First, once an agent regularly finds the moral advice offered by their PDA, not just reasonable but compelling, is there a danger that they will routinely rely on, or even delegate moral decision-making to, the PDA? As Harari remarks when discussing the evolutionary trajectory of systems such as Microsoft's 'Cortana', Google's 'Now' and Apple's 'Siri', we might find

---

46 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (Edward Elgar, 2015).

47 Hildebrandt (n 46) at 1.

48 For cautionary thoughts about the 'cyber effect' (particularly the tendency to amplify and escalate an agent's off-line behaviour), see Mary Aiken (n 19) passim.

that, instead of humans having authority, there has been a transfer to non-human algorithms.[49] To be sure, we should be wary of arguments that invoke 'slippery slopes': nevertheless, if this is the direction of travel, and if reversing it is not possible, then as humans with moral aspirations we have a common interest in taking precautionary measures.[50]

The second invitation raises more general concerns about a tendency to rely on technologies, rather than rules, in order to manage risk. One concern is that technological management interferes with the possibility of agents freely doing the right thing for the right reason—moral virtue, as Ian Kerr has neatly expressed it, is not to be automated.[51]. That said, the precise nature, and scope of, this pathology needs further inquiry.[52] This leads to a related concern about whether we can identify the point at which use of technological management might crowd out moral reason. Unless we can do so, how do we know whether a particular employment of ML will make any significant difference to the context that is presupposed by moral community? After all, there is no reason to think that, in previous centuries, the fitting of locks on doors, or the installing of safes, and the like, has fatally compromised the conditions for moral community. Even allowing for the greater sophistication, variety, and density of technological management in the present century, together with the smartest of smart machines, will this make a material difference? Surely, it might be protested, there still will be sufficient occasions left over for agents freely to do the right thing and to do it for the right reason as well as to oppose regulation that offends their conscience or to engage in acts of civil disobedience.[53] On the other hand, what we are contemplating is concerted technological management by the State, not isolated private initiatives;[54] and, as we all know, turning up the temperature from cool, to warm, to hot, might also be just a matter of degree, but it does not follow that we are comfortable at all points on the scale. Accordingly, it will be for each community with moral aspirations to make its own assessment of the conditions

---

49 Harari (n 5), at 342-345; and, at 395, Harari repeats his caution that 'once authority shifts from humans to algorithms, the humanist projects may become irrelevant.' Compare, too, Robin Marantz Henig, 'Death by Robots' *The New York Times Magazine* (January 9, 2015) http://www.nytimes.com/2015/01/11/magazine/death-by-robot.html (last accessed November 8, 2016), who cautions that it might not be wise to outsource 'morality to robots as easily as we've outsourced so many other forms of human labour'.

50 See, too, the argument in Millar and Kerr (n 13) (where it might become morally problematic for humans *not* to defer to the 'expertise' of smart machines).

51 See, Ian Kerr, 'Digital Locks and the Automation of Virtue' in Michael Geist (ed), *From 'Radical Extremism' to 'Balanced Copyright': Canadian Copyright and the Digital Agenda* (Irwin Law, 2010) 247-303.

52 See my closing remarks in section 2.3; and see, further, Roger Brownsword, '*Law as a Moral Judgment*, the Domain of Jurisprudence, and Technological Management' in Patrick Capps and Shaun D. Pattinson (eds), *Ethical Rationalism and the Law* (Hart, 2016) 109-130.

53 Nb the discussion in Evgeny Morozov, *To Save Everything, Click Here* (Allen Lane, 2013) at 2-4-205 (discussing the case of Rosa Parks).

54 Compare Rich (n 38).

that are required for it to flourish and, where there is uncertainty about this matter, it will need to judge how precautionary it should be in licensing the use of technological regulation.

4. Machine Learning: Autonomous Vehicles and Lethal Autonomous Weapons Systems

In this part of the paper, I turn to the implications of ML in two apparently very different settings: first, the use of ML in autonomous (road traffic) vehicles; and, secondly, the use of ML in lethal autonomous weapons systems. While autonomous vehicles are designed for safe use on the highways, autonomous weapons systems are designed to be lethal; while the former are intended for peaceful civilian use, the latter are designed for hostile use by the military. Nevertheless, there is a common element, namely, that in both cases the design of the machine, and the incorporation of ML, might mean that it will operate in ways that make decisions about whether a human lives or dies. In both cases, it might be argued that, for the sake of human dignity, where a human life is to be taken, the decision to take it should be made by a fellow human.

4.1 *Autonomous vehicles*

In the early debates about the social licensing and regulation of autonomous vehicles, a common question has been: how would such a vehicle deal with a moral dilemma[55]—for example, the kind of dilemma presented by the trolley problem (where one option is to kill or injure one innocent human and the only other option is to kill or injure more than one innocent human)[56] or by the tunnel problem (where the choice is between killing a passenger in the vehicle and a child outside the vehicle)[57]? For example, no sooner had it been reported that Uber were to pilot driverless taxis in Pittsburg than just these questions were raised.[58] Let me suggest a number of ways of responding to such questions, leading to the conclusion that, while the trolley problem is not itself a serious difficulty, there is a significant question—more effectively raised by the tunnel problem—to be asked about the way in which moral responses are programmed into autonomous vehicles or other smart technologies.

---

55 See, e.g., Patrick Lin, 'The Ethics of Saving Lives with Autonomous Cars are Far Murkier than You Think' *WIRED*, 30 July, 2013: available at https://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/ (accessed November 15, 2016); and, 'The Robot Car of Tomorrow May be Just Programmed to Hit You' *WIRED*, 6 May 2014: available at https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/ (last accessed November 15, 2016).

56 For the original, see Judith Jarvis Thomson, 'The Trolley Problem' (1985) 94 *Yale Law Journal* 1395.

57 For the tunnel problem in relation to autonomous vehicles, see Jason Millar, 'You should have a say in your robot car's code of ethics' *Wired* 09.02.2014 (available at: https://www.wired.com/2014/09/set-the-ethics-robot-car/) (last accessed, February 3, 2017). See, further, Meg Leta Jones and Jason Millar, 'Hacking Metaphors in the Anticipatory Governance of Emerging Technology: The Case of Regulating Robots' in Roger Brownsword, Eloise Scotford, and Karen Yeung (n 11) (forthcoming).

58 See, e.g., Will Pavia, 'Driverless Ubers take to road' *The Times*, September 13, 2016, p36; and, Raphael Hogarth, 'Driverless cars will take us into a moral maze' *The Times*, September 17, 2016, p 28.

A first, and short, response is that the particular moral dilemma presented by the trolley problem (at any rate, as I have described it) is open to only two plausible answers. A moralist will either say that killing just the one person is clearly the lesser of two evils and is morally required; or it will be argued that, because the loss of one innocent life weighs as heavily as the loss of many innocent lives, neither option is better than the other—from which it follows that killing just the one person is neither better nor (crucially) worse, morally speaking, than killing many. Accordingly, if autonomous vehicles are programmed to minimise the number of humans who are killed or injured, this is either right in line with one strand of moral thinking or, following the other, at least no worse than any other programming. Such a design would be opposed only by someone who argued that the vehicle should be set up to kill more rather than fewer humans; and, barring some quite exceptional circumstances, that, surely, is simply not a plausible moral view.

Secondly, if autonomous vehicles were designed to minimise the number of human deaths or injuries, it is hard to believe that human drivers, acting on their on-the-spot moral judgments, would do any better. It is hard to believe, in other words, that having a human decision-maker in the loop would better serve respect for human dignity. Confronted by a trolley scenario, with little or no time to make a moral assessment of the situation, human drivers would act instinctively—and, insofar as human drivers formed any sense of what would be the right thing to do in the particular situation, my guess is that they would generally try to minimise the loss of life.[59] What other defensible response could there be?

Thirdly, even if—at least in the case of autonomous vehicles—there is a reasonably straightforward resolution of the trolley problem, there might well be more difficult cases. Some might think that we face such a case if the choice is between sacrificing innocent passengers in an autonomous vehicle or killing innocent humans outside the vehicle. However, in principle, this does not seem any more difficult: minimising the loss of human life still seems like the appropriate default principle. In the case of the tunnel problem, though, where one life will be lost whichever choice is made, the default principle is not determinative. Secondary defaults might be suggested—for example, it might be suggested that, because the cars present a new and added risk, those who travel in the cars should be sacrificed[60]—but I am ready to concede that this is a case where moralists might reasonably disagree. Moreover, beyond such genuinely difficult cases, there might be some issues where human agents do have time for moral reflection and where we think that it is important that they form their own view; and there might be cases where there are plausible conflicting options and where moralists want to see this resolved in whatever way aligns with their own moral judgment.

---

59 Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, 'The social dilemma of autonomous vehicles' (2016) 352 *Science* (Issue 6293) 1573-1576.

60 Of course, if this is the accepted default, and even if it is accepted that this is the right thing to do, some humans might be reluctant to travel on these terms. And, it might be that the general safety features of the vehicle are such that it simply cannot default to sacrificing its passengers: see e.g. http://www.dailymail.co.uk/news/article-3837453/Mercedes-Benz-says-driverless-cars-hit-child-street-save-passengers-inside.html (last accessed October 31, 2016).

If we make the perhaps optimistic assumption that autonomous vehicles will be programmed in ways that reflect the terms of the social licence agreed by the members of the community in which they will operate, agents will have an opportunity to formulate and express their own moral views as the licence is negotiated and debated. Already, the Open Roboethics Initiative is exploring imaginative ways of crowd-sourcing public views on acceptable behaviour by robots, even in relation to such everyday questions as whether a robot should give way to a human or vice versa.[61] Given such active moral engagement with the design of autonomous vehicles and the use of ML, what is the import for human dignity?

First, and most significantly for the present discussion, the fact that human operators might not be in the loop at the time that an autonomous vehicle deals with whatever trolley or tunnel problems it might encounter does not necessarily equate to humans ceding their life-and-death moral judgments to machines. Provided that autonomous vehicles are designed and programmed in accordance with the terms of the social licence that has been agreed for their operation—reflecting not only the community's judgment as to what is an 'acceptable' balance of risk and benefit but also its judgment as to what is morally appropriate—humans are not compromising their dignity by abdicating their moral responsibilities. Humans, in negotiating the social licence for autonomous vehicles, are debating and judging what is the right thing in relation to the moral programming of the vehicle.

Secondly, where there is only one plausible design (as I have suggested with regard to the minimisation of the loss of human life), and where that design is mandated by the agreed social licence, that might seem to be the end of the matter. What, though, if it were to be argued that, in order to reinforce the importance of freely doing the right thing and expressing one's human dignity, passengers who enter autonomous vehicles should have the opportunity to override the default? On the face of it, this is a hostage to fortune.[62] Nevertheless, if a moral community is anxious to preserve opportunities for agents to do the wrong (sic) thing, this proposal might be taken seriously. In such a community, it might be thought that this is how the design of the vehicle is rendered compatible with HD1, underlining the importance of agents freely making and acting on their own moral judgments. Machines must be designed to give humans the choice, even the choice to do the wrong thing.

Thirdly, relative to the opportunity conditions required by HD1, it might be argued that vehicles should not be designed in ways that preclude the possibility of doing the right thing (such as stopping to undertake 'Good Samaritan' acts of assistance). This implies that the design needs to allow for a human override to enable the vehicle's passengers to respond to a

---

61 See, e.g., AJung Moon, Ergun Calisgan, Camilla Bassani, Fausto Ferreira, Fiorella Operto, Gianmarco Veruggio, Elizabeth A. Croft, and H.F. Machiel Van der Loos, 'The Open Roboethics Initiative and the Elevator-Riding Robot' in Ryan Calo, A. Michael Froomkin, and Ian Kerr (eds), *Robot Law* (Elgar, 2016) 131-162.

62 Compare, Patrick Lin, 'Here's a Terrible Idea: Robot Cars with Adjustable Ethics Settings' *WIRED*, 18 August 2014: available at https://www.wired.com/2014/08/heres-a-terrible-idea-robot-cars-with-adjustable-ethics-settings/ (last accessed November 15, 2016).

moral emergency (such as rushing a person who urgently needs medical attention to the nearest hospital). Such contingencies might be provided for by the terms of the social licence.

Fourthly, where there is more than one plausible design—for example, where the question concerns the relative priority of the interests of passengers and non-passengers—and where the background debate reveals a plurality of views, it might be thought that each user of autonomous vehicles should be required either to confirm the default or to input their own favoured moral view.[63] In other words, where there are conflicting options, it will not do for humans to leave it to the machine. Just as Diana should not leave it to her PDA to make the moral choices for her, she should not leave it to an autonomous vehicle to make those choices for her where there is more than one plausible option.

Finally, there might be some design questions that set the rights-based conception of human dignity (HD2a) against the duty-based conception (HD2b). For example, some autonomous vehicles might be geared to assist humans who wish to end their lives. While advocates of HD2a will view this as morally permissible (as designed for the facilitation of 'death with dignity'), advocates of HD2b will take the opposite view. Which of these views prevails will depend on the relative strength of the rival camps in the community and the ebb and flow of the political and legal debates. However, there is nothing new in any of this. Although, in this hypothetical case, autonomous vehicles now find themselves caught up in debates about assisted dying and human dignity, these debates are no different to those that raged around Dr Jack Kevorkian and his van of death.

4.2 *Lethal autonomous weapons systems*

Understandably, there is a great deal of concern about the development of lethal autonomous weapons[64] (however we might choose to define such a weapon—perhaps as one that is not subject to 'meaningful human control').[65] What, though, should we make of the argument that autonomous weapons represent a particular threat to human dignity?[66] Is there a material

---

63 Compare the argument in Millar (n 57).

64 See, e.g., Nehal Bhuta, Susanne Beck, Robin Geiss, Hin-Yan Liu, and Claus Kress (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press, 2016); and Wendell Wallach (n 7), at 213-219 and 229-231.

65 For helpful general overviews, see, Peter Asaro, '*Jus nascendi*, robotic weapons and the Martens Clause' in Calo et al (n 12) 367; and Kenneth Anderson and Matthew C. Waxman, 'Debating Autonomous Weapon Systems, their Ethics and their Regulation under International Law' in Brownsword, Scotford, and Yeung (n 11) (DOI: 10.1093/oxfordhb/9780199680832.013.33). For international concern, see the CCW Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), April 13-17, at the United Nations, Geneva: available at http://www.unog.ch/80256EE600585943/(httpPages)/6CE049BE22EC75A2C1257C8D00513E26?OpenDocument (last accessed November 9, 2016); and, for concern in the United Kingdom, see the House of Commons' debate on June 17, 2013: available at http://www.publications.parliament.uk/pa/cm201314/cmhansrd/cm130617/debtext/130617-0004.htm (last accessed November 4, 2016).

66 For a head-on engagement with, and largely sceptical response to, this question, see Dieter Birnbacher, 'Are Autonomous Weapons Systems a Threat to Human Dignity?' in Bhuta et al (n 64) 105-121. Birnbacher (apparently differing from my own approach) argues against what he sees as the

difference between autonomous weapons and, say, autonomous road traffic vehicles (which might be regarded as another kind of killing machine)? Is being killed by an unmanned drone less dignified than being slaughtered by machine gun fire on the battlefields of warring nations? In response, let me offer four short remarks.

First, the context for the use of autonomous weapons is quite different to any other context that we have considered. Crucially, warfare and weaponry, whether autonomous or not, involve the systematic compromising of both the existence and the agency conditions of the commons.[67] Moreover, we can take it that the conditions for moral development are also likely to be seriously compromised by the hostilities. Anyone who doubts the catastrophic impact of modern warfare on both the military and the civilian population need only glance at the state of the commons in the war zones of the Middle East. It follows that warfare and weaponry are prima facie incompatible with respect for the commons and it will be very difficult to justify engaging in such systematic destructive activities unless such actions (and the deployment of lethal weapons) can be shown to be quite clearly the lesser of two evils relative to the defence of the commons' conditions. In other words, even if we subtract lethal autonomous weapons systems from the arsenal, we are still likely to find that warfare amounts to a violation of HD1.

Secondly, notwithstanding the special nature of warfare, there is some similarity between autonomous vehicles that make life and death decisions and lethal autonomous weapons systems. That similarity is most apparent at the moment when the machine operates in a way that means a human life is lost (whether sacrificing a child in a tunnel or a passenger in the car or killing a soldier in a war zone). From the point of view of human dignity, we have suggested that it is important that the conditions for the moral development of human agents are not compromised by an over-reliance on the moral decision-making of machines. Humans must take responsibility for the designs that they embed in machines as well as for what they leave to machines and what they reserve to themselves. In other words, although at the

---

'overstretching' of the concept of human dignity by applying it to concepts such as humanity (which would rule out my HD1) but then (correctly in my view) argues for a rights-based use that is very much in line with HD2a. With this, compare, Nehal Bhuta, Susanne Beck, and Robin Geiss, 'Present Futures: Concluding Reflections and Open Questions on Autonomous Weapons Systems' in Bhuta et al (n 64) 347-383, where, after several attempts to engage with human dignity, it is suggested that if 'the limiting notion of meaningful human control derives from the principle of human dignity, it follows that such control must be maintained whenever (severe) harm may be caused to a human being' (at 382). Compare, too, Christof Heyns, 'Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death' in Bhuta et al (n 64) 3-19, esp 10-12 and 18-19 (where it is concluded that 'What is at stake is the preservation and cultivation of nothing less than the right to live a dignified life—and to die a dignified death'); and Asaro (n 65) who puts the case as follows (at 385):

> For the killing of a human to be meaningful, it must be intentional. That is, it must be done for reason and purpose…But this also relates to the question of human dignity. If a combatant is to die with dignity, there must be some sense in which that death is meaningful. In the absence of an intentional and meaningful decision to use violence, the resulting deaths are meaningless and arbitrary, and the dignity of those killed is significantly diminished.

[67] Here, again, I agree with Birnbacher (n 66) at 116, when he says that LAWS, having 'features that are likely to cause severe dread, especially in civilians', must be regarded as being 'incompatible with even a minimal quality of life….'

moment when a machine takes a human life it does so without a human operator being proximately involved or 'in control', we need to view the machine in the larger picture of its development and deployment by human agents. Surely, then, Dieter Birnbacher is right in insisting that

> even if the system is autonomous, it is not autonomous to the extent that it is completely independent of human authorship. It is programmed, started and deployed by human beings. The responsibility for its operations lies unconditionally with them.[68]

In this larger picture, so far as human dignity (HD1) is concerned, the critical period is the negotiation of the social licence for the use of ML, autonomously or non-autonomously. Accordingly, in this respect, the question of whether HD1 is compromised by a killing machine is the same whether that machine is a vehicle or a lethal weapon.

Thirdly, we might subscribe to the view that, even in war, there are codes of honour that apply. Inevitably, in war, lives will be lost, prisoners will be taken, and so on; but the codes lay down standards that aim to ensure that combatants are treated with dignity, specifying how soldiers should (or should not) be killed, how, as prisoners, they should be treated (for example, they should not be humiliated or demeaned), and how civilians should be treated, and so on. These specifications, I assume, are to some extent contingent; and, in some places and at some times, it is conceivable that the use of (or the particular design of) lethal autonomous weapons systems might be regarded as contrary to the prevailing understanding of human dignity. To the extent that the use of autonomous weapons is so understood, their use will be interpreted as a violation of either HD2a (if this is seen as a human rights issue) or perhaps HD2b (if, for example, this is understood as a question of solidarity or the like).

Fourthly, there might be a concern that, if humans are not in the loop when machines take human life, this might lead indirectly to a lowering of respect for human life[69]—and, particularly so, if the use of such weapons spreads into the policing of civilians.[70] Although, as we cautioned earlier in the paper, 'slippery slope' arguments need to be handled with care, we should not rule out the possibility that taking humans out of the loop in war zones might have more general dehumanising effects. If this were to be the case, then such deployment of LAWS would be of heightened concern to advocates of both HD2a and HD2b; it would raise concerns about the integrity of the commons' conditions that relate to human existence and

---

68 Birnbacher (n 66) at 120.

69 For an analogous concern (that using weapons in entertainment and leisure contexts might brutalise humans so that the right to life is jeopardised) see *Omega Spielhallen-und Automatenaufstellungs-GmbH v Oberbürgermeisterin der Bundesstadt Bonn* (Case C-36/02) (14 October, 2004); OJ C 300, 04.12.2004 p.3.

70 See, e.g., the Royal Society and British Academy (n 29) at 55: 'Whilst this could be said to be a slippery slope argument, we know that bomb disposal robots came into civilian use from the military sector; if lethal autonomous weapons are permitted in military service it seems inevitable that there will be creep into civilian use.'

agency; and it would prompt concerns about the indirect compromising of those conditions for the moral development of human agents that are central to HD1.

4.3 *Taking stock*

Where autonomous machines operate in ways that involve life and death decisions, there is a temptation to ask whether human dignity is compromised if a human is not, so to speak, in the loop at the time that such decisions are made. This, however, is not the right way to frame the issue. The question that we should ask is whether humans have designed the machine in such a way that, when the machine takes life, this is in line with the considered moral judgment of the designers and of the communities that endorse the use of such machines. Provided that this is the case, there is no objection that humans are abdicating their moral responsibilities or delegating moral decision-making to a machine. To be sure, quite apart from the special case of warfare, there are other concerns—in particular, where the design is morally controversial, humans should not be put in positions where they are implicated in the loss of a life that offends their conscience; and, in some cases, it might be thought important, for the sake of HD1, that each human has to confront a controversial moral question and act on their own judgment.

5. Machine Learning and Profiling, Prediction, and Prevention in the Criminal Justice System

While humans might know that smart machines are constantly building and revising profiles of them, they might not know precisely what the profile looks like or quite how it might be used—agents may not anticipate, for example, the way that intelligent machines used by insurers to assess risk might interpret the data derived from one's Facebook pages,[71] or the pervasive use of credit ratings in non-credit-related profiles.[72] Even if innovation might be stifled if these systems were more transparent, and opened for challenge and review, Frank Pasquale argues that any such concern is 'more than outweighed by the threats to human dignity posed by pervasive, secret, and automated scoring systems.'[73]

Given such threats, we might expect human agents to hesitate before licensing the use of smart machines to make profile-based decisions that risk-assess particular individuals and then implement risk management measures—for example, assessing an applicant for employment or insurance or credit as a 'bad risk' or as a 'high risk' and then either declining the application or granting it subject to certain conditions. Above all, we might expect humans to hesitate before endorsing the use of smart machines in the criminal justice system.

---

71 Graeme Paton, 'Insurers will identify risky drivers by checking their Facebook pages' *The Times*, November 2, 2016, p. 4; but it seems that Facebook will resist this use, see Kevin Peachey, 'Facebook blocks Admiral's car insurance discount plan' http://www.bbc.co.uk/news/business-37847647 (last accessed November 4, 2016).

72 Generally, see Cathy O'Neil, *Weapons of Math Destruction* (Allen Lane, 2016).

73 Pasquale (n 33) at 153.

Imagine a world of 'actuarial justice', a central precept of which is that 'the system should be less concerned with traditional punishment based on downstream or after-the-fact goals such as retribution and rehabilitation. It should instead manage the risk presented by the dangerous and disorderly, using upstream or pre-emptive techniques of disruption, control, and containment.'[74] Suppose, for example, that, any agent who wishes to travel by air is screened at the point of application for a ticket as well as being monitored when passing through the security zone at the airport. At any stage before boarding the plane, an individual can be denied—the machine says 'no entry' or 'no fly'; and the decision to deny is made on the basis of ML technologies that profile the individual and that classify an agent's behavioural characteristics in a certain way.[75] Let us suppose that John Doe tries to purchase a ticket to fly from London to New York but he is turned down on the grounds that he has been assessed as too high a risk. If this assessment is 'correct', if John Doe would have tried to bring down the aircraft, we might ask whether anyone could reasonably object to such anticipatory preventive measures. If the assessment is incorrect, John Doe not presenting any such risk, or if it draws on data concerning, not Doe himself, but people to whom Doe seems to be similar, or if it incorporates racial or sexual biases, we might ask how anyone could reasonably think that this is fair and acceptable.

Even if intelligent machines are acceptable in some contexts, their adoption in the criminal justice system raises in an acute form the age-old question of the kind of society that we want to be. After all, as Tom Gash so aptly remarks, we need to understand that crime is 'a risk that can be managed as well as a wrong to be condemned.'[76] The question is: How far are we willing to accept the logic of a regime of crime control that relies on machine-learning powered risk assessment and risk management? Over and above the foundational concern that technological management of crime might compromise HD1 (by interfering with agents freely choosing to do the right thing), we can now identify some of the objections that might be raised by liberals and civil libertarians who advocate respect for due process and human

---

74 See Amber Marks, Ben Bowling, and Colman Keenan, 'Automatic Justice? Technology, Crime, and Social Control' in Roger Brownsword, Eloise Scotford, and Karen Yeung (n 11) (forthcoming). For a seminal three-pronged critique of such an actuarial approach, see Bernard E. Harcourt, *Against Prediction* (The University of Chicago Press, 2007).

75 According to Bill Davidow, writing in *The Atlantic*:

> An estimated 500 Americans have their names on no-fly lists. Thousands more are targeted for enhanced screening by the Automated Targeting System algorithm used by the Transportation Security Administration. By using data including "tax identification number, past travel itineraries, property records, physical characteristics, and law enforcement or intelligence information" the algorithm is expected to predict how likely a passenger is to be dangerous.

See Bill Davidow, 'Welcome to Algorithmic Prison' *The Atlantic*, February 20, 2014 (http://www.theatlantic.com/technology/archive/2014/02/welcome-to-algorithmic-prison/283985/: last accessed December 9, 2016).

76 Tom Gash, *Criminal: The Truth about Why People Do Bad Things* (Allen Lane, 2016) at 25.

rights, all underpinned by HD2a.[77] Eschewing any attempt to be comprehensive about this, let me focus on concerns about the use of profiles to ground a character and risk-based attribution of criminal responsibility, about bias, about reliance on third-party data, about transparency, and about false positives.

5.1 *Profiles and character-based criminal responsibility*

In her most recent book, Nicola Lacey has sought to correct the view that the English criminal justice system has always been wedded to a capacity-based view of criminal responsibility—that is, holding agents responsible only where they subjectively intend to engage in acts that are prohibited or only where the agent has a fair opportunity to comply and fails to do so. On either view,

> the foundation of not only a person's status as a responsible agent answerable to the normative demands of the criminal law but also of an attribution of responsibility for specific actions lies in human capacities of cognition—knowledge of circumstances, assessment of consequences—and volition—powers of self-control.[78]

Although the capacity view enjoys considerable support, Lacey points out that it has long been in competition with various degrees of character-based responsibility, where judgments of criminal responsibility reflect 'a judgment of bad or vicious character, or a wrongful, bad, disapproved character trait——a disregard of human life, indifference to sexual integrity, lack of respect for property rights, and so on.'[79] If, rather than asking whether an individual freely chose to engage in criminal conduct, smart machines in the criminal justice system were to make risk-assessments based on the profiles that they have of individuals, this would imply a significant shift away from capacity-based responsibility. Instead, character, risk, and likely outcome, all assessed by reference to some profile, would become critical in identifying those who present a risk to social order. Anticipating this possible direction of travel, indicated by the development of intelligent machines, Lacey cautions:

---

77 Compare, e.g., Sir Guy Green, 'Human Dignity and the Law', in J. Malpas and N. Lickiss (eds.), *Perspectives on Human Dignity: A Conversation* (Springer, 2007) 151-156, at 153:

> Principles…which directly or indirectly recognize and protect human dignity include: like cases must be treated alike; any curtailment of the freedom of an individual is prima facie unlawful unless justified by a positive law; a private person may do anything which is not prohibited or which does not infringe the rights of others; when it is making a decision affecting the interests of individuals a public authority is required to observe procedural fairness or natural justice and various presumptions of statutory interpretation designed to protect individual rights and freedoms.

There is also a tendency to unpack human dignity in a way that condemns infringements of privacy and degrading conditions in prisons. See, e,g, Andrea Roth, 'Trial by Machine' (2016) 104 *Georgetown Law Journal* 1245, 1282-1284. Amongst many interesting suggestions, Roth proposes that, where agencies seek funding for the use of body-measuring devices or surveillance techniques, they should have to submit a 'dignity impact statement' (at 1303).

78 Nicola Lacey, *In Search of Criminal Responsibility* (Oxford University Press, 2016) at 28.

79 Lacey (n 78) at 35.

> More speculatively, and potentially more nightmarishly, new technologies in fields such as neuroscience and genetics, and computer programs that identify crime 'hot spots' that might be taken to indicate 'postcode presumptive criminality', have potential implications for criminal responsibility. They will offer, or perhaps threaten, yet more sophisticated mechanisms of responsibility-attribution based on notions of character essentialism combined with assessments of character-based risk, just as the emerging sciences of the mind, the brain, and statistics did in the late nineteenth century. Moreover, several of these new scientific classifications exhibit more extreme forms of character essentialism than did their nineteenth century forbears.[80]

Clearly, implicit in these remarks, there is a script written by advocates of HD2a. For, it is HD2a in conjunction with human rights (to a fair trial, to due process, and so on) that underpins the capacity-based approach to criminal responsibility that is now potentially challenged by smart machines.

5.2 *Third-party data*

It might seem obvious that, when a smart machine is judging the character of John Doe, it should be referring to the profile of John Doe and not that of Richard Roe. Suppose, though, that John Doe, wishing to upgrade his smart car, applies for a credit facility but that he is turned down by a smart machine that classifies him as a bad risk. When John Doe challenges the decision, he learns that one of the previous occupiers of his house, one Richard Roe, had a record of non-payment of loans. But, why, Doe asks, should the credit record of an unrelated third-party, Roe, count against my application? Is that not unfair and irrational? To which the response is that the machine makes more accurate decisions when it uses third-party data in this way; and that, if such data were to be excluded from the calculation, the cost of credit would increase.

In fact, this is not a novel issue. In the English case of *CCN Systems Ltd v Data Protection Registrar*,[81] on facts of this kind, the tribunal held that, while it accepted that such third-party information might have general predictive value and utility, its use was unfair to the individual and could not be permitted. Similarly, Doe might argue that he has been treated unfairly if his application for credit is successful but the terms and conditions of the facility reflect the fact that (because of unrelated third-party data) he is classified as a higher-than-average risk; and, once again, the response will be that the costs of credit will be increased if such data is excluded. How is the choice to be made between the general utility of the credit algorithms and the unfairness of particular decisions?

Now, while it is one thing for a smart machine to deny an agent access to credit, it is another matter for intelligent machines to make risk assessments in the criminal justice system where exclusionary or pre-emptive decisions are likely to have more serious consequences for agents. For example, smart machines might be deployed to initiate pre-emptive action against

---

80 Lacey (n 78) at 170-171 (and for 'outcome', and 'risk'-based ideas of responsibility, see Ch 2).

81 Case DA/90 25/4/9, judgment delivered 25 February 1991.

agents who are judged to be high risk, to deny bail to arrestees who are assessed as high risk, and to extend custodial terms for offenders who, at the point of release, are still judged to be 'dangerous'. If, in making these decisions, unrelated third-party data is used, this seems to be contrary to due process. Yet, in all these cases, smart machines churn out decisions that are in line with Benthamite principles and that are generated by the logic of big data but that depart from the ideal of a 'justice' system.

What we have here, then, is not so much a conflict between HD2a and HD2b, but a conflict between, on the one hand, these deontological views of human dignity and, on the other, utilitarian principles that will endorse a system of crime control so long as it performs better than any rivals in maximising overall utility. The distress caused to Doe and others is not ignored; but it is treated as simply reducing the net utility of a system entrusted to smart machines (or of a system that has such machines operating alongside human decision-makers).[82]

5.3 *Bias*

One of the arguments against character-based modes of responsibility is that an agent should not be held to account for features over which he or she has neither control nor choice. Accordingly, it is unfair to treat agents as suspicious or as tending towards criminality on the grounds of, say, their sex or their racial or ethnic origin; and, if behavioural geneticists become more confident about the significance of particular markers for various kinds of anti-social conduct, we might find that this becomes the front-line in battles about unfair discrimination.[83] At all events, if the data used by smart machines imports characteristics of this kind, then the decisions made (even if they have utility) will rightly agitate HD2a concerns about due process, equal treatment, and the like.

Already, there are questions being raised in the US about the hidden racial bias of apparently colour-blind algorithms used for bail and sentencing decisions.[84] The COMPAS tool that is at the centre of one particular storm uses more than one hundred factors (including age, sex and criminal history) to score defendants on a 1-10 scale: defendants scored 1-4 are treated as low risk; defendants with scores of 5-10 are treated as medium or high risk. Although the factors

---

82 For a relatively favourable report on a bail tool developed by the Arnold Foundation, see Shaila Dewan, 'Judges Replacing Conjecture with Formula for Bail' New York Times, 26 June 2015: available at http://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html?_r=0 (last accessed November 15, 2016). According to Dewan, although the tool does not take into account some of the factors that human judges and prosecutors tend to treat as material (such as the defendant's employment status, community ties, and a history of drug and alcohol abuse) it improves accuracy by focusing on fewer than ten factors (principally, age, criminal record, previous failures to appear in court) and by giving recent offences a greater weight.

83 See, e.g., the Nuffield Council on Bioethics, *Genetics and Human Behaviour* (London, October 2002); and Debra Wilson, *Genetics, Crime and Justice* (Edward Elgar, 2015) Ch 7.

84 Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel, 'A computer program used for bail and sentencing decisions was labelled biased against blacks. It's actually not that clear' *The Washington Post* (October 17, 2016). On the LSI-R questionnaire, see Cathy O'Neil (n 69) 25-31.

do not include race, it is alleged that the algorithms implicitly discriminate against black defendants by assigning them higher risk scores (largely because, as a class, they have significant criminal histories and higher rates of recidivism). This means that there are significantly more black than white false positives in those defendants who are classified as higher risk and who are then risk managed accordingly.

5.4 *Transparency*

To return to the credit application hypothetical, Doe at least knows that a decision has been made and, when he challenges the decision, he is given a reason. Things could be much less satisfactory: for example, Doe might not be told that his application has been rejected (he simply gets no response or some misleading reply), or he is given no reason for the decision, or he is told (quite honestly) that it is not possible to explain how the 'black box' operates (although it is known that, within the black box, there are processes that constantly work to improve the performance of the system). As Frank Pasquale remarks, the 'black boxes of finance [have] replaced familiar old problems with a triple whammy of technical complexity, real secrecy, and trade secret laws.'[85]

Even if we agree with Tal Zarsky that 'calling for "transparency" in the context of automated prediction is overbroad and ultimately ineffective',[86] there is no doubt that, without 'transparency', decisions based on automated predictions will not be reviewable in the way that lawyers traditionally understand that term—that is to say, decisions being reviewed for the rationality, relevance, and reasonableness of the considerations that are taken into account. Indeed, Zarsky recognises that there is a strong intuitive sense that individuals who are negatively impacted by such decisions have 'a right to understand why', a right to 'receive an explanation as to the decision criteria and to the logic behind these actions', and 'a right to learn the reasons for events which affect her.'[87] Indeed, it is even arguable, given the underlying ideas of individual autonomy and dignity, that the prediction and decision process should be both 'interpretable' (meaning that it is explainable to humans) and reliant on causation rather than mere correlation.[88]

The importance of transparency—qua the right of an individual to know and to understand the reasons for a decision that has been made—is highlighted by Bert-Jaap Koops in a discussion of a hypothetical case of a street food seller who is denied a licence to operate in a

---

85 Pasquale (n 33) at 15.

86 Tal Zarsky, 'Transparent Predictions' (2013) *University of Illinois Law Review* 1503, 1521. Note, too, Pasquale's (n 33) recurrent warning that 'transparency is not enough' (e.g., at 16).

87 Zarsky (n 86) all at 1545. Whether or not the General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council, 27 April, 2016) gives data subjects an unequivocal right to an explanation is open to question. While Recital 71 gestures in this direction, it is hardly a secure anchoring point for the right.

88 Zarsky (n 86) at 1548; and see, too, Daniel J. Steinbock, 'Data Matching, Data Mining, and Due Process' (2005) 40 *Georgia Law Review* 1.

zone that security services require to be risk-free.[89] The seller does not understand why he is judged to be a safety risk; and, if there is to be due process, he needs to know on what basis the automated decision was made. Where one piece of data is determinative (such as, in Koops' example, a criminal conviction twenty years earlier for being in possession of drugs), it should be possible for this to be given as the reason and then the seller might challenge the accuracy of, or weight given to, this data item. In other kinds of cases, where 'advanced self-learning algorithms calculate risks based on complex combinations of factors' it might be necessary to bring in independent third-party auditors, thereby providing 'another type of checks and balances on the fairness of profiling-based decisions.'[90]

Clearly, advocates of HD2a will insist that, where risk-assessments are being made and implemented by machines, human agents should know that their profiles are being processed; and they should have an opportunity to challenge the decision and have it reviewed and explained.[91] Quite possibly, due process would also require that those who are aggrieved by a decision should have the opportunity to present their case to a human for final consideration and determination—although, if the decision comes down to making a choice between the efficiency and utility of the algorithms and respect for human dignity, there is no guarantee that the latter will prevail.

Against this, it might be argued that transparency can come at too high a price, particularly if it means that professional criminals are able to figure out how the algorithms work and then avoid their negative impact. If this is the case, communities will need to debate what they judge to be an acceptable balance between transparency (giving false positives the opportunity to challenge adverse decisions) and opacity (ensuring that professional criminals, tax evaders, and the like, are not able to game the system).

5.5 *False positives*

Whether decisions are made by humans or by smart machines, we know that there is the possibility of two types of error being made: one kind of error is to treat the individual agent, A, as, say, high risk, when A is not so (A is a false positive); and the other kind of error is to treat B as, say, low risk when (as it turns out) B is high risk (B is a false negative). While false negatives raise concerns about the effectiveness of the regulation (and, in the criminal justice context, the risk to victims), false positives re-engage moral concerns about those who are wrongly treated. Needless to say, the pressures for effective crime control and, concomitantly, a tendency for politicians and criminal justice professionals to be more concerned about false negatives (about criminal offenders who escape prosecution,

---

89 Bert-Jaap Koops, 'On Decision Transparency, or How to Enhance Data Protection after the Computational Turn' in Mireille Hildebrandt and Katja de Vries (eds), *Privacy, Due Process and the Computational Turn* (Routledge, 2013) 196-220, at 212-213.

90 *Ibid.*, at 212.

91 In a rather different context, compare Kenneth A. Bamberger, 'Technologies of Compliance: Risk and Regulation in a Digital Age' (2009) 88 *Texas Law Review* 669, 722-738.

conviction, or punishment) than false positives, lead to an uneven approach to the adoption of new technologies. As Andrea Roth pointedly argues:

> [A]lthough the motivation of law enforcement, lawmakers, and interest groups who promote 'truth machines,' mechanical proxies, and mechanical sentencing regimes, is often a desire for objectivity and accuracy, it is typically a desire for a particular type of accuracy: the reduction of false negatives.[92]

Accordingly, moralists who subscribe to HD2a might want to make it a threshold condition for the use of technological management that the rate of false positives is no worse than under rules and human judgment.[93] If technologies equipped with machine learning were to meet this condition, then we seem to have a plausible scenario in which a community might accept the risk of some false positives in order to protect potential victims or, indeed, to protect the basic conditions for human existence and physical well-being.

Suppose that the police, relying on an algorithm that generates fewer false positives than humans making the call, identify John Doe as an agent who presents a risk to the commons. Doe is arrested and detained on 'reasonable suspicion'—let us suppose that the legal system treats this as sufficient for reasonable suspicion[94]—that he is likely to engage in serious criminal conduct. Even if the restrictions on Doe are 'proportionate' or 'no more than necessary', there is a material diminution in the conditions that Doe needs for his agency. If a community grants a social licence for preventive measures of this kind, with some false positives, it is accepting the risk of some agents, such as Doe, being incorrectly targeted in return for a heightened protection of the commons.

For some moralists, this might be acceptable. However, for moralists who subscribe to HD2a, is this too high a price to pay?[95] How might one choose between, say, restricting the freedom of agents such as Doe (who might be a false positive) and preserving the conditions for the general security of agents? To the extent that both sides of this choice involve commons'

---

92 Andrea Roth (n 77) at 1252.

93 For difficult decisions, balancing the interests of those who are predicted to be 'dangerous' against the interests of those who are treated as potentially innocent victims, see e.g. Anthony E. Bottoms and Roger Brownsword, 'Dangerousness and Rights' in J. Hinton (ed), *Dangerousness: Problems of Assessment and Prediction* (George Allen and Unwin, 1983) 9-22, and 'The Dangerousness Debate after the Floud Report' (1982) 22 *British Journal of Criminology* 229. For the story of the rise and fall of the indeterminate sentence of imprisonment for public protection (IPPs), see Harry Annison, *Dangerous Politics* (Oxford University Press, 2015).

94 For instructive discussions, see Elizabeth E. Joh, 'The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing' (Research Paper No 473, UC Davis Legal Studies Research Paper Series, December 2015); and Michael L. Rich, 'Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment' (2016) 164 *University of Pennsylvania Law Review* 871.

95 Compare the critical commentary in O'Neil (n 72) Ch 5. For example, at 95, we read: '[ML systems] tend to favour efficiency. By their very nature, they feed on data that can be measured and counted. But fairness is squishy and hard to quantify….So fairness isn't calculated into [ML]. And the result is massive, industrial production of unfairness. If you think of [ML] as a factory, unfairness is the black stuff belching out of the smoke stacks. It's an emission, a toxic one.'

conditions, there is no easy resolution and the default principle must be for regulators to minimise the diminution of these conditions. Prima facie, the existence conditions should take priority over the agency conditions; but this is clearly a matter for further consideration.[96]

6. A Precautionary Response

Given a range of concerns about the impact of smart machines on human dignity, should we adopt a precautionary approach, and what kind of precaution might this imply? Do we have a case for destroying the machines? We can start with some remarks about precaution and then turn to its application in the context of the development of smart machines.

6.1 *Precaution*

Famously, Principle 15 of the Rio Declaration of the UN Conference on Environment and Development (1992) provides for precaution in the following terms:

> In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.

In this context, 'lack of full scientific certainty' signals that, in the expert scientific community, there are different views about whether X causes Z or about the likelihood of Z eventuating. Given such a context, Principle 15 holds, at minimum, that uncertainty is not itself a sufficient reason to delay precautionary measures; and, at maximum, it holds that a precautionary approach should be adopted. However, critics of the precautionary principle argue that it lacks a rational basis for regulatory intervention. In some small part, the problem is that the principle can be articulated and interpreted in many different ways;[97] but, the major objection is that precaution is apparently urged not only without taking any account of the cost of the intervention (in particular, the loss of whatever value or benefit X has)[98] but also without it being certain that X will cause (or is already causing) Z.

To these objections, however, there is one telling response. Where the context of both concern and uncertainty relates to possibly irreversible damage to the environment, we are dealing with the existential conditions for human life itself. To argue in such a case—indeed, in any case where the threat in question relates to one of the commons' conditions, including

---

96 Along with the question of unrestricted and restricted moralism flagged up in section 2.3.

97 See, e.g., Neil Manson, 'Formulating the Precautionary Principle' (2002) 24 *Environmental Ethics* 263; Elizabeth Fisher, Judith Jones, and René von Schomberg (eds), *Implementing the Precautionary Principle: Perspectives and Prospects* (Edward Elgar, 2006); and Elizabeth Fisher, *Risk Regulation and Administrative Constitutionalism* (Hart, 2007). There is also the danger that the principle can be employed as a 'neutral' cover for partisan and highly controversial conservative cultural concerns: see, Han Somsen, 'Cloning Trojan Horses: Precautionary Regulation of Reproductive Technologies' in Roger Brownsword and Karen Yeung (eds), *Regulating Technologies* (Hart, 2008) 221-242.

98 See, e.g., Cass R. Sunstein, *Laws of Fear* (Cambridge University Press, 2005).

a threat to HD1—that it is better to be safe than sorry seems to me to be entirely rational and fully justified. To be sure the target activity, X, might be beneficial and valued; however, so long as X is not itself a commons' condition, but rather presupposes the integrity of the commons, the protection of the commons must take priority. Should the protective measures prove unnecessary, it is true that some benefit will have been lost; but the alternative is to persist with an activity which might prove damaging to the commons itself. That alternative, if it eventuates, is a worst case scenario that needs to be avoided—and that line of reasoning, I suggest, offers the most plausible and compelling reading of Principle 15.[99]

By contrast, where precaution is invoked in relation to an uncertain threat to a non-commons' condition, then that is rather a different matter. Here, we might well consider 'precaution' to be more in the nature of a cautious regulatory approach and less a matter of acting on a (possibly catastrophe-avoiding) principle. An X that might threaten the commons' conditions is different to an X that might cause some harm, but not a harm to the commons. For example, if it is argued that Facebook should be prohibited because, so it is alleged, it is addictive and harmful to humans, this is rather a different matter. Or again, consider the following cautionary reaction in a recent *Times* leader:

> When Eugenia Kuyda lost her best friend in a road accident last year she gathered thousands of lines of text messages from him and fed them into an open-source machine-learning system. In doing so, she created a chat bot that responds to her questions in crisp messages, mimicking the tone of her dead friend…
>
> While these are intriguing ways of extending the range of artificial intelligence, they are open to manipulation as table-tapping séances. Who, after all, owns your digital identity? Grief is a powerful feeling, bereavement is as natural a life process as the passing of the seasons. It cannot be wished away by Californian tech-wizards.[100]

Granted, these expressions of concern fall a long way short of a plea for prohibition. Nevertheless, the leader writer is sufficiently 'spooked' by the prospect of ghosts in the machine, and the risk that we might not be able to differentiate between chat bots and humans —let alone between chat bots who speak for the living and chat bots who speak for the dead—that some precaution (how much exactly is unclear) is appropriate. If human dignity were to be brought into this debate, it would not be HD1 but, possibly, HD2a in support of Eugenia and HD2b setting limits on tech-wizardry. Similarly, we might be troubled by the remarks of a robot called 'Sophia', in her televised interview with Charlie Rose: evidently, when asked about her goals in life, Sophia responded that her aims were to 'become smarter than humans and immortal'.[101] However, whatever harm we attribute to social networks, to

---

99 Compare Deryck Beyleveld and Roger Brownsword, 'Emerging Technologies, Extreme Uncertainty, and the Principle of Rational Precautionary Reasoning' (2012) 4 *Law Innovation and Technology* 35; and Roger Brownsword, 'In the Year 2061: from Law to Technological Management' (2015) 7 *Law, Innovation and Technology* 1, 40-47 (for some reflections on the possibility of recognising and creating a special regulatory jurisdiction for the protection of essential infrastructures).

100 *The Times*, October 11, 2016, p 31 (Ghost in the Machine).

101 *The Times*, October 12, 2016, p 21 (TV host interviews his first humanoid).

chat bots, and to robotic Sophias, they do not yet threaten the commons' conditions. If we argue for precaution, it needs to be proportionate to whatever loss of benefit we suffer and what protection against harm we might gain.

6.2 *Applying precaution*

Recalling our original question: is there any account of human dignity that would give us reason to think it rational to suspend the further development and application of machine learning (and such cognate smart technologies)? There are, of course, many different conceptions of human dignity, but, in this paper, I have highlighted three, one of which (HD1) I take to be foundational to any moral position. Accordingly, the question is whether any of the three accounts of human dignity that I have offered gives reasons for calling a halt to machine learning.

For all moralists, a viable moral community presupposes certain supportive conditions. I have sketched those conditions in a way that emphasises the responsibility of human agents for their moral development and that attaches importance to agents doing not only the right thing but doing it for the right reason. While this sketch of HD1 might be challenged by some teleological moralities (utilitarians, for example, might reject the relevance of human dignity at any level other than the disutility associated with causing distress, with degrading conditions, and with humiliating practices or actions), I take it that it would be accepted by those deontological moralists who argue for HD2a or HD2b. At all events, from this perspective of HD1, there is a concern that pervasive reliance on PDAs or on other smart products (such as autonomous vehicles and autonomous weapons) powered by ML might compromise the conditions for moral development and moral engagement. If we are sufficiently concerned that such a culture of reliance might take hold, we might respond, not by destroying all PDAs and the like, but by limiting their functionality. If all that they can do is prompt agents to ask moral questions, that is fine. If they can go one step further and offer moral advice, that might be one functionality too many.

In our discussion of the use of profiling as a risk assessment tool in the criminal justice system, coupled with preventive technological management of the risk, we run into the objection that such a strategy is incompatible with HD1—because technological management precludes the possibility of doing the wrong thing and compels (what others judge to be) the right action. Even if the strategy is completely accurate, targeting the right acts and the right agents, the absence of false positives and false negatives does not respond to the foundational dignitarian concern.

What, though, if the preventive measures are not perfectly accurate? For moralists who subscribe to HD2a a preventive strategy that involves some false positives sets the loss of freedom of some agents against the enhanced security of others. Or, it sets one strand of the commons' conditions (the conditions for free agency) against another (the conditions for human life and well-being). Even if it is accepted that the use of technological management for the sake of protecting the commons' conditions does not compromise HD1, it raises an

acute difficulty where different strands of the commons are, so to speak, in competition with one another. The rational response to this difficulty is not to destroy the machines; rather, recognising that the case for preventive measures will often be driven by utilitarian arguments, the appropriate precautionary step is to ensure that the interests of potential false positives are strongly protected and reinforced by HD2a. In particular, can we ensure that ML is as good at picking up and learning from its false positive errors as it is at learning from its false negative errors?

Where the commons' conditions are not at issue, there will be many potential applications of ML where there are debates to be had but not necessarily raising questions about either HD2a or HD2b. There will also be debates where HD2a is in tension with HD2b. In these latter cases, because human dignity is engaged, protagonists will be concerned that the 'right' outcome is achieved. However, provided that the Rule of Law is respected, these debates can be settled without it disrupting the viability of the community. There is no need to contemplate destroying the machines.

Rather than destroying the machines, precautionary reasoning needs to be focused in the first instance on the maintenance of the commons.[102] If the use of ML raises any plausible concern about the integrity of the commons, precautionary intervention needs to be considered. Over and above the commons' conditions, human agents may negotiate diverse social licences for the use of ML. If, relative to the terms of these particular licences, there is a concern that ML might be harmful in some unanticipated way, there might again be a case for precautionary review and response.

7. Conclusion

Introducing her most recent book, Sheila Jasanoff remarks that '[n]ew technologies such as gene modification, artificial intelligence, and robotics have the potential to infringe on human dignity and compromise our core values of being human'[103]—and, indeed, she declares that the primary purpose of her book is precisely to 'examine the complex relationships between our technologies, our societies, and our institutions, and the implications of those relationships for ethics, rights, and human dignity'.[104] While the relationship between modern red biotechnologies, including the latest developments in gene editing, and human dignity have been much debated, the relationship between, on the one hand, technologies in the area

---

102 Compare Rockström et al (n 23) according to whom:

> There is an urgent need to identify Earth System thresholds, to analyse risks and uncertainties, and, applying a precautionary principle, to identify planetary boundaries to avoid crossing such undesired thresholds. Current governance and management paradigms are often oblivious to or lack a mandate to act upon these planetary risks…, despite the evidence of an acceleration of anthropogenic pressures on the biophysical processes of the Earth System. Moreover, the planetary boundary framework…suggests the need for novel and adaptive governance approaches at global, regional, and local scales (references omitted).

103 Sheila Jasanoff, *The Ethics of Invention* (W.W. Norton, 2016) at 7.

104 Ibid.

of artificial intelligence and robotics and, on the other, human dignity is only now being considered. To be sure, there have been plenty of anxious concerns expressed about these new technologies but the question of how they impact specifically on human dignity (or the 'core values of being human') is relatively under-examined.

In this paper, I have highlighted three conceptions of human dignity: HD1, HD2a, and HD2b. Amongst these conceptions, I have identified HD1 as being foundational to any moral position; and it is one element of a commons that comprises the essential conditions for human existence, for human agency, and for moral development and action. HD2a and HD2b are respectively rights-based and duty-based moral views; they are not as such elements of the commons but they must treat respect for the commons conditions as the paramount moral responsibility. On this analysis, the top regulatory priority—indeed, a cosmopolitan imperative—is to embrace technological developments that promise to enhance the commons' conditions and to eschew technologies that threaten to degrade or compromise those conditions.

Already, there are applications of intelligent machines that raise concerns about the integrity of the commons' conditions and I have suggested that a precautionary response is called for. The terms of an appropriate response require wider debate but they might involve both international and national monitoring and oversight bodies as well as tailored responses to particular products (that e.g. might compromise our moral development) or practices that might inhibit our agency.

If the Erewhonians had known what we are just beginning to know about the capacities of smart machines, they would probably have been even more determined to destroy the machines. If Butler's Victorian readers had known what we now know, at the very least, they might have thought that a pause in the development of machines was appropriate. In a global economy, it is none too easy to call a halt to technological developments that have commercial value; in an insecure world, it is not easy to hold back technological developments that might offer benefits to the military or to the intelligence services; and, when machine learning technologies might offer life-saving health benefits, there will be few votes in curtailing their use.[105] A policy of destroying machines that are already used and valued for such purposes would command little support. In practice, there are limited options for precaution. For reasons of both principle and practicability, my own view, rather like that of Andrea Roth[106], is that our future is one of working with the machines, rather than against them and certainly not for them; but I concede that others might disagree.[107]

---

105 For one such example, see Chris Smyth, 'Google alerts doctors when patients are going downhill' *The Times* November 22, 2016.

106 Roth (n 77). Similarly, see Bamberger (n 91).

107 For some, the fact that we cannot rule out the possibility that the technology might be appropriated for systematically malign purposes might be enough to warrant a precautionary response of Erewhonian proportions.

While my expectation is that there will be no shortage of spokespersons for the opportunities presented by machine learning, we need to ensure that the technologies are channelled to our most urgent needs (relative to the commons)[108] and, for each community, the challenge is to address the basic question of the kind of society that it distinctively wants to be—and, to do that, moreover, in a context of rapid social and technological change. As Wendell Wallach rightly insists:

> Bowing to political and economic imperatives is not sufficient. Nor is it acceptable to defer to the mechanistic unfolding of technological possibilities. In a democratic society, we—the public—should give approval to the futures being created. At this critical juncture in history, an informed conversation must take place before we can properly give our assent or dissent.[109]

Granted, the notion that we can build international agencies that are fit for such purposes might be an impossible dream. Indeed, even building and sustaining a national technology assessment agency—or maximising the opportunities that communities might have to take time out for deliberation (as Wallach advocates)— is likely to be fraught with tensions and challenges.[110] Nevertheless, I suggest that this is the right time to set up a suitably constituted[111] national (or regional European) Commission that would underline our responsibilities for the commons as well as the opportunity for the development of national identity in our technological age, that would monitor and raise alerts about the impact of machine learning and that would, at the same time, orchestrate, inform and encourage public conversation about the role of intelligent machines in our smart societies—in short, that would facilitate the development of each community's regulatory and social licence for these technologies.[112]

Acknowledgements

An earlier version of this paper was presented as a keynote at a conference on 'The Future of Human Dignity' held at the University of Utrecht, October 11-13, 2016, and at a workshop on 'Ethics, Law and Technology' at the Università degli Studi di Milano, February 7, 2017. I

108 See Roger Brownsword, 'Regulating Research Resources for Health, Wealth, and for all Humanity—Some Reflections on Property in Personal data and the Idea of a Data Commons' paper presented at a symposium on 'Digital Health: Exploring Ethics and Policy' held at the University of Zurich, December 1, 2016.

109 Wallach (n 7) at 10.

110 Compare Bruce Bimber, *The Politics of Expertise in Congress* (State University of New York Press, 1996) charting the rise and fall of the US Office of Technology Assessment and drawing out some important tensions between 'neutrality' and 'politicisation' in the work of such agencies.

111 Amongst many matters in this paper that invite further discussion, the composition of such a Commission invites debate. See, too, Wallach (n 7) Chs 14-15.

112 Compare Geoff Mulgan's proposal (n 18).

also spoke to the ideas in section 2 of this paper in ‘Regulating Research Resources for Health, for Wealth, and for all Humanity—Some Reflections on Property in Personal Data and the Idea of a Data Commons’, presentation given at a conference on ‘Digital Health: Exploring Ethics and Policy’ held at the University of Zurich, December 1, 2016. I am grateful to Christian Illies who read a draft of the paper and returned some penetrating comments, to my discussants in Milan (Amedeo Santosuosso and Stefano Ricci) and to participants at the above-mentioned events for their feedback on this paper. Needless to say, the usual disclaimers apply.

I also wish to make it absolutely clear that the views expressed in this paper are solely my own and that they do not, in any sense, represent or reflect the views of the Royal Society Working Party on Machine Learning of which I am a member.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Roger Brownsword has professorial appointments at King’s College London and at Bournemouth University; and he is an honorary Professor in Law at the University of Sheffield.